Language and country preponderance trends in MEDLINE and its causes

By Alvar Loria, PhG
aloria@quetzal.innsz.mx
Scientific Researcher

Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán
Vasco de Quiroga #15
14000 Mexico City
Mexico

Pedro Arroyo, MD, MSc
parroyo@funsalud.org.mx
Scientific Consultant

Fundación Mexicana para la Salud
Periférico Sur 4809
14610 Mexico City
Mexico

Objective: The authors characterized the output of MEDLINE papers by language and country of publication during a thirty-four-year time period.

Methods: We classified MEDLINE's journal articles by country of publication (Anglos/Non-Anglos) and language (English/Non-English) for the years 1966 and from 1970 to 2000 at five-year intervals. Eight English-speaking countries were considered Anglos. Linear regression analysis of number of papers versus time was performed.

Results: The global number of papers increased linearly at a rate of 8,142 papers per year. Anglo and English papers also increased linearly (6,740 and 9,199, respectively). Journals of Non-Anglo countries accounted for 25% of the English language increase (2,438 per year). Only Non-English papers decreased at a rate of 1,056 fewer papers per year. These trends have led to overwhelming shares of English and Anglo papers in MEDLINE. In 2000, 68% of all papers were published in the 8 Anglo countries and 90% were written in English.

Conclusions: The Anglo and English preponderances appear to be a consequence of at least two phenomena: (1) editorial policy changes in MEDLINE and in some journals from Non-Anglo countries and (2) factors affecting Non-Anglo researchers in the third world (publication constraints, migration, and undersupport). These are tentative conclusions that need confirmation.

INTRODUCTION

A growing dominance of English as the ''lingua franca'' of research communications has been reported by Garfield for the Science Citation Index (SCI) [1]. The overall trends of language have seldom been documented for MEDLINE [2], despite MEDLINE's importance as a source of information for meta-analysis research [3]. In a previous paper dealing with clinical and basic research tendencies in MEDLINE [4], the authors observed that language was associated with ba-sic or clinical research in the mother-child health care area. This observation led us to explore the trends of language and country of publication regardless of area of research. Our objectives were to characterize the language trends in MEDLINE, especially of English in journals published in countries that do not have English as a mother tongue. For our purposes, we considered Anglo publications those published in eight countries (three divisions of the United Kingdom, Australia, Canada, Ireland, New Zealand, and the United States) and as Non-Anglo those published in all other countries.

J Med Libr Assoc 93(3) July 2005

381

Table 1
Number and percentage of MEDLINE articles by language and country of publication

Year	Total MEDLINE journal articles	English	%	Non-English	%	Anglo journals	%	Non-Anglo journals	%
1966	174,400	93,173	53	81,227	47	76,066	44	98,334	56
1970	213,066	125,496	59	87,570	41	98,663	46	114,403	54
1975	243,118	163,388	67	79,730	33	123,573	51	119,545	49
1980	258,329	185,536	72	72,793	28	137,870	53	120,459	47
1985	307,866	233,853	76	74,013	24	168,703	55	139,163	45
1990	367,568	293,265	80	74,303	20	214,027	58	153,541	42
1995	389,170	340,261	87	48,909	13	255,502	66	133,668	34
2000	468,191	419,108	90	49,083	10	317,705	68	150,486	32

MATERIAL AND METHODS

MEDLINE search

The search was done on Monday, August 11, 2003, using the MEDLINE database of the PubMed system <http://www.ncbi.nlm.nih.gov/PubMed/>. We selected a Monday for the search because MEDLINE is continually updated from Tuesday to Saturday. We performed our search for specific years beginning with 1966, when MEDLINE's database originally began, and every five years from 1970 to 2000. We restricted our search to publications classified by MEDLINE as ''Journal-article'' to prevent the presence of non-articles such as letters, editorials, interviews, commentaries, newspaper articles, and others. Non-articles have been increasing in MEDLINE (i.e., from less than 0.4% of all papers from 1966 to 1975, to 5% in 1980 and 1985, 7% in 1990, and 10% in 1995 and 2000).

All journal articles were classified by language (English and Non-English) using the MEDLINE language (LA) field and by country of publication (Anglos and Non-Anglos) using the country of publication (CP) field. Eight English-speaking countries were considered Anglos (Australia, Canada, England, Ireland, New Zealand, Scotland, United States, and Wales) and all others Non-Anglos.

Statistical analysis

Linear regression analysis (LRA) was used to explore the relationships between yearly number of papers (Y values) versus year of publication of the papers (X values). LRA is very commonly used if a linear relationship exists between variables, as the slope will allow the prediction of future Y values corresponding to future X values (e.g., stock market behaviors are frequently analyzed by LRA).

The regression parameters of an LRA are three: the slope, the intercept, and the correlation coefficient. For our purposes, the slope is the most important one and represents the change in Y (papers/year) due to a change in one unit of X (1 year). If the relationship is linear, then the slope represents the mean change that has been occurring, in our case, in the number of papers every year throughout the time period 1966 to 2000. Steeper slopes indicate a larger change (a larger increase if the slope is positive, a larger decrease if negative).

The intercept is the value of Y when X has a value of zero. In our case, we use the year 1966 as the zero value, so that our intercepts are the LRA-predicted number of papers in 1966. The correlation coefficient is used here as an index of goodness of fit of the XY data to a straight line. A coefficient of +0.90 (or a negative −0.90) implies that there is almost a perfect fit of the data to a straight line (square of ±1 = 1 = 100%).

RESULTS

Table 1 shows the raw data of all MEDLINE journal articles published in the time period of 1966 to 2000 and classified separately by language and country of publication. Globally, the yearly number of papers in MEDLINE has nearly tripled in the time period (468,191/174,400 = 2.7), and the English and Anglo contributions have more than quadrupled (419,108/93,173 = 4.5 and 317,705/76,066 = 4.2, respectively), whereas Non-English papers have decreased 40% (49,083/81,227 = 0.60). The number of papers from Non-Anglo journals have increased at a rate (150,486/98,334 = 1.5) nearly 3 times lower than the 4.2 increase in papers from Anglo countries. In the year 2000, these trends led to overwhelming preponderances of English (90% of all MEDLINE papers) and Anglo country of publication (2 of every 3 papers).

Table 2 gives the number of papers cross-classified by language and country of publication. The journals from Non-Anglo countries have also contributed to the preponderance of English, as they have steadily increased their English papers in the time period (101,650/17,972 = 5.7). This increase has been at the expense of Non-English papers, which increased from 1966 to 1970 and then steadily decreased, reaching record lows in 1995 and 2000.

The Non-English papers in Anglo journals, mostly Canadian papers in French, have also been decreasing. Their number has mostly been below 1% of the Anglo production, and, thus, we performed our subsequent analysis without considering language in the journals from Anglo countries.

The linear regression parameters (see ''Statistical Analysis'') are shown in Table 3. There is a good fit to a straight line for all changes (correlations near to plus/minus one). The overall rate of increase was 8,142 papers per year, contributed mostly by Anglo papers (83% of the increase) and less by Non-Anglos (17% of

Table 2
Number and percentage of MEDLINE articles cross-classified by language and country of publication

	Non-Anglo					Anglo				
Year	Total journal articles	Non-English	%	English	%	Total journal articles	Non-English	%	English	%
1966	98,334	80,362	82	17,972	18	76,066	865	1	75,201	99
1970	114,403	86,565	76	27,838	24	98,663	1,005	1	97,658	99
1975	119,545	79,201	66	40,344	34	123,573	529	—	123,044	100
1980	120,459	71,908	60	48,551	40	137,870	885	1	136,985	99
1985	139,163	73,484	53	65,679	47	168,703	529	—	168,174	100
1990	153,541	73,954	48	79,587	52	214,027	349	—	213,678	100
1995	133,668	48,591	36	85,077	64	255,502	318	—	255,184	100
2000	150,486	48,836	32	101,650	68	317,705	247	—	317,458	100

Anglo countries = Australia, Canada, England, Ireland, New Zealand, Scotland, United States, Wales.

the increase). English had the largest rate of increase (9,199 papers/year), whereas the category of Non-English papers was the only one showing a decrease of 1,035 fewer papers per year. In contrast, the journals from Non-Anglo countries increased their English papers and contributed 25% of the global increase in English during the 34-year time period (2,438/9,199 = 27%).

DISCUSSION

Our analysis shows that English has displaced all other languages in MEDLINE. This trend started in the 1970s, and, since then, Non-English has been decreasing, with large drops in 1980 and in 1995, especially. It is too early to know whether Non-English will level off at this new level of about 10% of all MEDLINE articles (Table 1). Our observations, based on 5-year intervals, were very similar to those of Sousa Escandon et al. [2], who used all years from 1966 to 1999. They found the same 40% drop of Non-English papers and that 89% of all papers in 1999 were in English, similar to our finding of 90% in all 2000 papers. On the other hand, these authors do not analyze their information using country of publication.

It seems clear that Non-Anglo journals have contributed to the preponderance of English in MEDLINE, but their shift to English appears to have contributed to their smaller share of papers in MEDLINE. We believe their contribution to MEDLINE will continue to decrease due to several factors, some related to MED-

LINE itself and others due to differences in research capabilities of Anglo and Non-Anglo countries or to Third World journal editors.

MEDLINE editorial policies

MEDLINE has been increasingly deselecting journals from Non-Anglo countries and may have developed a reluctance to include new journals that publish partly in Non-English. To illustrate these comments, we use the inclusion of Mexican journals in MEDLINE. As shown in Table 4, thirty Mexican journals were included in MEDLINE during the time period 1966 to 2000, and twenty of them have been deselected and have not reentered MEDLINE despite the fact that all accept papers in English. Deselection of Mexican journals appears to have increased in the last twenty years: eleven journals were deselected and only two new journals added in the time period 1980 to 2000.

We believe that MEDLINE's deselection of journals from Non-English countries was due to a need to confront the global increase of scientific papers, clearly exemplified here by the number of MEDLINE papers going from less than 200,000 per year in 1966 to nearly half a million in 2000 (Table 1) with an ominous straight line of increase that shows no sign of declining. MEDLINE furthermore appears to have been less drastic in its deselection than SCI, which abruptly lowered its selected journals from more than 4,000 to less than 3,000 in 1978. We think the SCI deselection was mostly of journals from Non-English countries (the

Table 3
Linear regression parameters of the number of articles shown in Tables 1 and 2

	Correlation coefficient	Slope papers/year	Calculated values		Ratio 2000/1966
			1966	2000	
All papers	0.99	8,142	167,348	444,186	2.7
English global	0.99	9,199	78,832	391,587	5.0
Anglos global	0.98	6,740	61,966	291,116	4.7
Non-Anglos global	0.89	1,403	105,382	153,070	1.5
Non-Anglos in English	1.00	2,438	17,807	100,696	5.7
Non-Anglos Non-English	−0.88	−1,035	87,575	52,373	0.6

Anglo countries = Australia, Canada, England, Ireland, New Zealand, Scotland, United States, Wales. The calculated value of 1966 corresponds to the intercept of the regression.

Table 4
Changes in the number of Mexican journals in MEDLINE from 1966 to 2000

	Added	Deselected	In MEDLINE
1966	19	0	19
1970	6	7	18
1975	3	2	19
1980	1	5	15
1985	1	3	13
1990	0	0	13
1995	0	2	11
2000	0	1	10*
Totals	30	20	

* One journal became an Anglo publication in 2000.

five Mexican journals still remaining in SCI at that time were excluded). In our view, the SCI action appears to be an administrative convenience decision in which language was very probably involved. It is difficult to assume that such a massive deselection could be based on a judgment of poor scientific quality in the deselected journals.

Editorial policies of journals of Non-Anglo countries

We feel that some editors have contributed to the preponderance of English by accepting papers only in English. In 1990, one of the ten Mexican journals remaining in MEDLINE in 2000 (Table 4) changed to only English and anglicized its name to Archives of Medical Research (formerly Archivos de Investigación Médica de México). This change has also been the case in other countries, especially those that do not use the Roman alphabet.

In addition, the anglicized Mexican journal was classified by MEDLINE in 2000 as an Anglo country publication (produced in the United States by Elsevier who promotes and sells it worldwide). This tempting strategy contributes to an increase in the scientific and commercial visibility of any journal, be it Anglo or Non-Anglo, and may become a factor further favoring the preponderance of Anglo publications.

Publication constraints on researchers of Non-Anglo countries

Researchers from less-developed countries are being pressured to increase their output in English and to submit it to journals published in Anglo countries, at least in Latin America as suggested by the paper by Bunout and Reyes [5]. Another example is provided in our country by the National Council of Research and Technology of the Federal Government (Spanish acronym CONACYT), the main supporter of research in Mexico since it started operations in 1984. In its yearly evaluation of the output of its researchers, CONACYT specifically asks for the exclusion of any information from papers published in local journals, which is the same as telling researchers to publish only in English and in foreign, developed countries. We know personal instances of Mexican researchers who have given up

trying to publish after being rejected by journals in Anglo countries, a rejection that not infrequently is more on the basis of form (bad English) than substance. This loss of information, which will not reach even the so-called gray literature, is sad.

To our knowledge, there is no information regarding the possible repercussions of publication constraint policies such as CONACYT's (i.e., whether there has been an increasing participation of Non-Anglo researchers in publications from Anglo countries). A possible way of finding this information is MEDLINE's address field of the institution of the first author (AD), but the retrieval process for country of the institution proved to be lengthy and incomplete. The AD field is free-format text with variable information concerning the adscription of the paper's first author that may include department, unit, institution, city, country, and email. Not infrequently, it has no information and, more frequently, lacks one or more of the other pieces of information, including country. In addition, we were unsuccessful in automatically retrieving the name of the country from the AD field. It would be useful if MEDLINE could add a country field for the first author in addition to its current AD field.

Research capability

This factor inevitably has repercussions on the scientific output of any country. The global output has increased despite the poor research infrastructure in less-developed countries, whose situation appears to be worsening. The largest Latin American countries are now assigning less than 0.5% of their gross national products to research and development. Nevertheless, dwindling resources have not led to the disappearance of existing journals, although they may have affected the emergence of new journals and, thus, the Non-Anglo share of MEDLINE papers.

On the other hand, the Anglo preponderance is not due only to lagging scientific production in Third World countries. Non-Anglo developed countries are participating in two manners that appear to be increasing with time: more papers in English as discussed above (Table 2) and scientific migration, such as the current migration of scientific researchers from Europe to the United States [6].

The language trends presented here have led to the advisability of using other databases such as EMBASE for European literature and Lilac for Latin American papers in meta-analytic studies [7]. However, the problem of identifying papers from researchers in Third World countries in Africa, Asia, and even in Latin America who are forced to write in English is a difficult one.

Studies aimed at establishing the quality of Non-English papers show that research published in Non-English is on a par with English publications, at least in randomized clinical trials [8], and that Non-English journals have less publication bias than journals in English [9]. These observations have led Petitti [3] to declare that the practice of limiting meta-analysis to studies published in English must be condemned.

CONCLUSIONS

Our LRA of the number of articles in MEDLINE showed an increase at a rate of 8,142 papers per year from 1966 to 2000. Linear increases were also observed for Anglo and English papers (6,740 and 9,199 papers/year, respectively). These trends have led to overwhelming shares of English and Anglo papers in MEDLINE, so that, in 2000, 68% of all papers were published in Anglo countries and 90% written in English.

Other interesting findings were that 25% of the increase of English papers came from journals published in Non-Anglo countries (rate of increase of 2,438 papers/year) and that the preponderance of Anglo and English papers was associated with lower numbers of Non-English papers, which decreased at a rate of 1,056 papers per year.

Our tentative conclusion was that several factors are contributing to these trends, mainly the editorial policies of MEDLINE and of some journals published in Non-Anglo countries, to greater research capabilities of Anglo countries than Non-Anglo countries, and, finally, to constraints imposed on researchers in Non-Anglo countries. An additional factor might be the recent brain drain of European researchers migrating to the United States.

ACKNOWLEDGMENTS

REFERENCES

1. GARFIELD E. The diverse roles of citation indexes in scientific research. Rev Invest Clin 1998 Nov–Dec;50(6):497–504.

2. SOUSA ESCANDON MA, GONZALEZ GUITIAN C, GONZALEZ FERNANDEZ MM. ¿Qué lengua hablará MEDLINE en el próximo milenio? [Which language will MEDLINE speak in the next millenium?]. Arch Esp Urol (Spain) 2000 Mar;53(2):93–9.

3. PETITTI DB. Meta-analysis, decision analysis, and cost-effectiveness analysis. New York, NY: Oxford University Press, 2000:55–6.

4. LORIA A, ARROYO P. Research trends in mother-child healthcare, 1966–1995. Matern Child Health J (USA) 2000 Mar;4(1):59–66.

5. BUNOUT D, REYES H. Manuscritos biomédicos escritos por autores chilenos y publicados en revistas internacionales en 1997. [Biomedical papers written by Chilean authors published in international journals in 1997]. Rev Med Chil 1998 Jun;126(6):677–88.

6. Plugging Europe's brain drain. Time (European edition) 2004 Jan 19;163(3):32–9.

7. SUTTON AJ, ABRAMS K, JONES DR, SHELDON TA, SONG F. Methods for metanalysis in medical research. New York, NY: Wiley, 2000.

8. MOHER D, FORTIN P, JADAD AR, JUNI P, KLASSEN T, LE-LORIER J, LIBERATI A, LINDE K, PENNA A. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet 1996 Feb 10;347(8998):363–6.

9. EGGER M, ZELLWEGER-ZAHNER T, SCHNEIDER M, JUNKER C, LENGELER C, ANTES G. Language bias in randomised controlled trials published in English and German. Lancet 1997 Aug 2;350(9074):326–9.